Honors Thesis: Supplementary Material

LCC 4700: Undergraduate Thesis Writing

"The Evolutionary Impact of Functional RNA Secondary Structures within Protein-Coding Regions in Yeast"

Charles Warden

Advisor: Dr. Soojin Yi

Reviewers: Dr. Soojin Yi and Dr. King Jordan

Literature Review

Previous use of duplicate genes produced by whole genome duplication in yeast

Because analysis of gene duplicates provides some insight into determining if the predicted fRNAs from this study are genuine, it is useful to review the significance of gene duplication and previous studies investigating the evolutionary patterns resulting from a whole genome duplication in yeast. Gene (and genome) duplication is a very important source of novel genetic information (Ohno 1970, Lynch and Conery 2000). Following gene duplication, the duplicate pair can undergo 1) nonfunctionalization – one duplicate becomes nonfunctional (typically via accumulation of mutations), 2) subfunctionalization – the functions of the predecessor gene are divided amongst the duplicates or 3) neofunctionalization – one of the duplicates evolves a novel function. Assuming both pairs are conserved, the duplicates may show asymmetric divergence, which means that one duplicate accumulates substitutions in the nucleotide sequence more quickly than the other duplicate. This phenomenon can be explained if the more slowly evolving duplicate is under greater functional constraint. However, the more rapidly evolving duplicate can *either* be under lesser constraint than the other duplicate *or* undergoing rapid change due to adaptive benefits produced in the duplicate gene. This study will examine how functional RNA (fRNA) structures differ in duplicate genes created by a whole genome duplication (WGD) in yeast (Kellis et al. 2004). The purpose of this project is to gain some preliminary data that may illuminate the role of RNA secondary structures in gene regulation and expression and how genomic duplication affects these important biological processes.

The large number of duplicate gene pairs in post-WGD species of yeast has already proven useful in comparative studies investigating functional asymmetry of gene duplicates (see Kim and Yi 2006). There has also been a recent study to investigate differences in the evolution of duplicates produced by the whole genome duplication versus duplicates produced by smaller scale duplications (SSD) in S*accharomyces cerevisiae* (Guan et al. 2006). One reason that this study may be useful for my research project the list of WGD and SSD paralogs used in this study may be useful in comparing the proportion of fRNA secondary structures found in and around duplicate versus non-duplicate genes. However, synonymous sites show saturation in S*accharomyces cerevisiae* (Guan et al. 2006), and there are other epistemological problems, such as long-branch attraction, with dating ancient events using phylogenetic analysis of molecular data from related yeast species(Fares et al. 2006). There may be methods to theoretically estimate the age of the smaller scale duplications; for example, there has been a previous study that used a maximum likelihood method to estimate the age of the whole genome duplication in yeast (Sugino and Innan 2005). However, sufficiently rigorous analysis of the distribution of fRNA secondary structures in and around SSD pairs is probably beyond the scope of what can be accomplished with this project, so the distinction between WGD and SSD pairs is probably most useful in the broad sense in that I will need to understand that there are limits to the extent to which I can apply my findings to gene duplication in general.

Another recent study indicated that there is widespread neofunctionalization that occurred recently after the whole genome duplication in yeast (Bryne and Wolfe 2006). The authors base this conclusion on the observation that the genes that are homologous to

a faster evolving paralog in post-WGD species also appear to evolve faster than their respective paralogs (implying that the asymmetric evolution probably occurred before speciation); furthermore, the authors provide evidence that the faster evolving duplicate is also rarely essential, and this is also fits the model for neofunctionalization (Bryne and Wolfe 2006). The authors also cite an earlier paper (Lynch and Force 2000) to show that subfunctionalization is theoretically unlikely to be an important mechanism for the retention of gene duplicates; however, it should be noted that this paper also discusses situations where subfunctionalization would be relatively *more* common (Lynch and Force 2000). This is important for my study because subfunctionalization (and perhaps neofunctionalization) requires biological modularity of function, and I hope that functional RNAs *both* within and around coding regions can help identify possible mechanisms for a modular basis of asymmetric evolution in gene duplicates (in protein coding and regulatory regions, respectively).

Network Analysis of Gene Duplication

There have been a number of previous studies to investigate the importance of gene duplication in regards to genomic evolution (Lynch and Conery 2000). There have also been some structural studies to investigate the role of gene duplication on protein domains and/or regulatory networks. One relatively early study showed that gene duplication can create a significant increase in the number of transcription factor interactions, which regulate gene expression (Teichmann and Babu 2004). A later paper demonstrated that singletons had a disproportionately high number of connections, meaning that duplicate pairs of genes with high connectivity were typically not retained (Hughes and Friedman 2005). This study also showed that connections tended to be

partitioned amongst gene duplicates, which is consistent with the model of subfunctionalization (Hughes and Friedman 2005). A more recent study combined network analysis with structural information about the location of protein-protein interactions to show that proteins with many interactions at the same site on a given protein will be more likely to gain additional interaction partners following gene duplication than a protein that would have to evolve a novel binding interface (Kim et al. 2006).

In fact, there is a very recent paper describing widespread neofunctionalization shortly after the whole genome duplication in yeast on co-expression networks for *Saccharomyces cerevisiae* (Conant and Wolfe 2006). Such structural studies are important because they can help provide a mechanistic view of how gene duplication affects the molecular interactions that govern gene regulation and expression. In this investigation, I will test for correlations between the number of functional RNAs in and around duplicate genes in yeast and connectivity in protein-protein interaction networks, so it will be useful for me to understand previous studies of the gene duplication using network analysis.

Role of functional RNAs as non-coding RNAs (cont.)

MicroRNAs (miRNAs) play a role in silencing the expression of specific protein-coding genes that correspond to very specific mRNAs (Pickford and Cogoni 20003). It has been observed that miRNA-mediated regulation has an impact on coding sequence in mammals due to the complementary property of antisense-binding sites in coding regions (Hurst 2005). It will be especially interesting to look for this evolutionary pattern in yeast because, to the best of my knowledge, no microRNAs have been discovered in

S*accharomyces cerevisiae*, although dsRNA has been found in other species like *S pombe* (Ouellet et al. 2006). Thus, a search for miRNAs in *S. cerevisiae* would result in the discovery of novel fRNAs and help explain selection in nearby coding regions (which would be in duplicate genes in the case of this study).

Role of functional RNAs in messenger RNAs (cont.)

Furthermore, this project may help illuminate sources of codon usage bias in yeast. A recent study found that a subset of approximately 60 duplicate gene pairs produced by the whole genome duplication in yeast show decelerated evolution and strong codon usage bias (Lin et al. 2006). In fact, this paper proposes that codon usage bias was more important than gene conversion for causing a decelerated rate of evolution (as was previously proposed in Kellis et al. 2004). So, it will be interesting to see if this investigation can provide any evidence to support either of these two opposing hypotheses.

Results

Asymmetric Evolution of Gene Duplicates (cont.)

Another interesting observation is that there a negative correlation between SRR and SRK (Signed Relative Connectivity in protein-protein interaction networks) within coding regions (albeit at a p-value <0.25) while there is no significant correlation between SRR and either SRA (Signed Relative protein Abundance) or SRF (Signed Relative Fitness). This observation may also relate to added constraint within coding regions because conserved RNA secondary structures within protein-coding regions cause a reduction in the number of possible interaction sites that can be produced for a given protein structure. This observation is also interesting because previous studies

have proposed that functional correlations with connectivity are mostly determined by other functional measures (Drummond et al. 2006). So, our results may indicate that connectivity may be a functional measure that is more biologically significant than originally conceived.

Functional Enrichment of Genes Containing fRNAs (cont.)

It is also interesting that there are a significant number of protein-coding genes that interact with non-coding RNA, and this may relate to the enrichment with protein biosynthesis genes (because such genes often interact with rRNAs and/or tRNAs). It is also interesting that proteins associated with fRNA secondary structures within their protein coding regions are often associated with protein biosynthesis because it is well known that duplicate genes retained from the whole genome duplication in yeast are also enriched with genes responsible for protein biosynthesis (Guan et al. 2006). This enrichment of WGD pair may be due to a fundamental difference between duplicate pairs produced by the whole genome duplication in yeast and smaller scale duplications (SSD): for example, WGD pairs may be part of a duplicated *system* that requires preservation of specific protein and/or RNA interactions while SSD pair can only duplicate *parts* of biological systems and tweak the connectivity of their respective system. If this hypothesis regarding gene duplication is true, then this may allay concerns of bias towards conserved genes in the dataset.

Methods

Multi-Species Alignment: Preliminary Search

Before EvoFold can produce predictions of fRNA folds, genomic sequences must be pre-aligned and screened to produce small fragments for EvoFold to analyze because

EvoFold cannot handle genomic sequences much greater then 750 basepairs in length (Pedersen et al. 2006). Conserved blocks of a seven species alignment of yeast species were selected using data created by the phastCons program (Siepel et al. 2005). Only phastCons blocks containing regions of synteny described in the supplementary material from Kellis et al. 2004 were selected. Initial and Final genes were determined from the file "Matches_by_chromosome_with_syn.xls" in the supplementary material from Kellis et al. 2004. Coordinates for initial and final genes were determined as described with the "*Saccharomyces cerevisae* Gene Annotation" section in the supplementary material.

Preliminary analysis of various types of multi-species alignments was based upon EvoFold predictions in 20 regions synteny (see Table S1). During this initial screening process, four (*Saccharomyces cerevisae*, *S. paradoxus*, *S. mikatae*, and *S. baynus*), five, or six species alignments (with *Saccharomyces kudriavzevii* and/or *Saccharomyces castelii*) were extracted from the phastCons alignment. Realignment of these nucleotide sequences with CLUSTALW was also tested in order to evaluate the accuracy of the alignment (Thompson et al. 1994). See Table S3 for a more detailed description of the various folds found within two duplicate regions of synteny.

Multi-Species Alignment: Stringent Dataset

EvoFold predictions from 4-,5-, and 6- species comparisons were created from 1) all phastCons blocks within any region of synteny, as described by Kellis et al. 2004, and 2) the underlying Multiz alignment for the set of WGD pairs, as described by Kellis et al. 2004 (see Figure S9). For the Multiz alignment, the goal was to obtain the sequence 300 nucleotides before the translation start site and ending 300 nucleotides after the translation stop site. In the event that a Multiz block did not begin or end within 300

nucleotides of the coding sequence, the flanking region was extended to the beginning or end of, respectively, the first or last Multiz block. The sequences for this entire region were joined together and sliding window analysis was conducted as described in the preliminary search section. The reason we tested EvoFold predictions from the Multiz alignment was that we found that the phastCons blocks could split known non-coding RNAs into multiple blocks (thus resulting in loss of recovery for that particular non-coding RNA), so we wanted to try and ensure that no folds were missed because of the locations of the phastCons blocks. However, we were not able to recover all of the predictions made from the phastCons blocks in these regions and there were very few known annotations that could be used to gauge the accuracy of the various levels of predictions, so we decided not to use these predictions in our final dataset. If the Multiz folds were less accurate than the phastCons folds, the reason may be that some of the windows for the Multiz folds would include sequences with both high and low sequence similarity and it may be difficult for EvoFold to determine that the window as a whole contains a significant signal for the presence of a functional RNA.

EvoFold predictions from the phastCons blocks were than screened using the RNAz program (Washietl et al. 2005 – see RNAz Program section for more details). These two programs make predictions in fundamentally different ways (see Figure S11) and correlations for SR measures are very different for every dataset (see Figure S5-S9). It was very difficult to objectively determine the optimal method to screen the original set of EvoFold predictions. However, the optimal screening process was determined by comparing the proportion of known annotations recovered for a particular method relative to the number of folds retained by imposing a more "strict" significance level

(although it was unclear if the FPS values for the EvoFold program really corresponded to more accurate predictions – see EvoFold Program section for more details). For example, increasing the RNAz p-value from 0.5 to 0.9 reduced the total number of folds to 38% of the original total number of folds but still retained 73% percent of known tRNAs, 60% of snoRNAs, 60% of snRNAs, 100% of rRNAs, and 93% of miscellaneous RNAs predicted by the RNAz program for the 5-species alignment at the p-value of 0.5. In the end, we decided that the most accurate dataset would most likely be the set of EvoFold predictions produced by the 5-species alignment (with an FPS value greater than 0) that were independently verified by the RNAz predictions made using the 6-species alignment with an RNAz p-value of 0.99. An EvoFold prediction was considered to be independently verified by the RNAz program if the middle of the EvoFold prediction was within an RNAz prediction (so that at there would be at least 50% overlap between the two predictions). Furthermore, there did not seem to be any RNAz predictions that were shorter than 10 nucleotides, so we assumed that the vast majority of EvoFold predictions that were less than 10 nucleotides were not likely to form a stable RNA secondary structure. Thus, all EvoFold predictions used in the stringent dataset were also required to be greater than 10 nucleotides in length.

EvoFold Program

The EvoFold program was used to predict fRNA secondary structures in post-WGD species of yeast, and it took approximately one month to compete a whole genome screening with four species alignment (*Saccharomyces cerevisae*, *S. paradoxus*, *S. mikatae*, and *S. baynus*), as described in the "Multi-Species Alignment" section. Using a Linux cluster, all other comparisons took a little over a week.

The significance of the fold level was determined by a folding potential score (FPS). FPS is a length normalized likelihood-ratio score and is defined as follows: FPS= log ($P(x|\phi_{fRNA})/P(x|\phi_{bg})$) / *l*, where $P(x|\phi_{fRNA})$ refers to the probability that a sequence fits an fRNA structural model, $P(x|\phi_{bg})$ refers to the probability that the sequence fits the background model (i.e. no-fRNA structure model), and *l* refers to the length of the fold (defined by the outermost basepair of a fRNA structure) (Pedersen et al. 2006, supplementary material). Interestingly, there is no significant correlation between the length of a fold and FPS in the 4-species (ρ=0.014, p-value=0.6138; see Figure S6), or 5-species (ρ=0.001, p-value=0.9716; see Figure S7) comparisons. However, there is a significant correlation between FPS and length for the 6-species comparison (ρ=0.051, p-value=0.034; see Figure S8). This correlation would be expected because longer folds are supposed to contain a higher proportion of significant folds (Pedersen et al. 2006). Furthermore, it would be expected that the FPS score should show a stronger correlation when more species are used; although this is true for the 6-species comparison, there is no such increase in significance level for the 5-species comparison. However, it is possible the calculations used in this study may not utilize the proper values for $P(x|\phi_{fRNA})$ and/or $P(x|\phi_{bg})$ if these values chance depending upon the length of the window that EvoFold uses for the whole genome screening (i.e. if the score varies with the length of flanking sequence around a given fold, then it will be difficult to determine the optimal size without already knowing the size of each individual fold).

RNAz Program

Unlike the EvoFold program, the RNAz program relies mostly on thermodynamic information to predict RNA secondary structures (Washietl et al. 2005). However, this

program also utilizes the same 4-, 5-, and 6- species alignments from the phastCons blocks that were used for the EvoFold predictions, and the program gives a "bonus" decrease in minimum free energy (MFE) of covariance is observed between each of the yeast sequences and the consensus sequence (and poorly conserved secondary structures are likewise given an increase in the minimum in the overall MFE value). RNAz gives a p-value for its predictions, but this p-value is not equivalent to the traditional statistical definition of a p-value; however, the proportion of retained known annotations was always larger than the proportion of folds remaining after increasing the p-value, so it is probably safe to assume that this RNAz p-value does indicate the accuracy of the dataset consistently. As described in the "Multi-Species Alignment: Stringent Dataset" section, it seems clear that the RNAz program is able to be more accurate when more species are used in the alignment and the RNAz p-value is increased (see Table S4 and Figure S10 and S11).

***Saccharomyces cerevisae* Gene Annotation**

Whenever possible, coordinates for genes were based upon the "SGD Genes" table for the *Saccharomyces cerevisae* genome sequence available on the UCSC Genome Browser (Karolchik et al. 2003). Unless otherwise noted in Table 2, all coordinates for genes came from the "SGD Genes" table. If a gene name could not be found within this table, data available from the "SGD Other" table from the UCSC Genome Browser was used (Karochik et al. 2003). In the event that the gene annotation could still not be found, coordinates from the SGD_features.tab file available from the FTP for the *Saccharomyces* Genome Database were used for that particular gene (Cherry et al. 1997).

Duplicate genes in post-WGD species were determined using the source code from the DupGenes.html file in the supplementary material from Kellis et al. 2004.

Fold Annotation

Fold location was determined by the position of the middle of each fold (i.e. a fold was in a particular category if >50% of the fold was in that type of region). All folds were categorized as coding, intronic, or intergenic. When searching for folds around duplicate genes, folds were categorized as coding, intronic, 5` flank, or 3` flank. In order for a fold to be considered in a 5` or 3` flanking region of a duplicate gene, the middle of the fold had to be within 300 nucleotides of, respectively, the start or stop site of the ORF for that gene. If 5` or 3` flanking regions contain an adjacent ORF, then the name of this ORF was noted in the table of results for each of those screenings.

Conclusion / Future Analysis

Asymmetric Evolution of Gene Duplicates

Another interesting observation is that there a negative correlation between SRR and SRK (Signed Relative Connectivity in protein-protein interaction networks) within coding regions (albeit at a p-value <0.25) while there is no significant correlation between SRR and either SRA (Signed Relative protein Abundance) or SRF (Signed Relative Fitness). This observation may also relate to added constraint within coding regions because conserved RNA secondary structures within protein-coding regions cause a reduction in the number of possible interaction sites that can be produced for a given protein structure. This observation is also interesting because previous studies have proposed that functional correlations with connectivity are mostly determined by other functional measures (Drummond et al. 2006). So, our results may indicate that

connectivity may be a functional measure that is more biologically significant than originally conceived.

Molecular Mimicry

Molecular mimicry is a phenomenon where protein structure mimics RNA structure (or, potentially, vice versa). One of the best known examples of molecular mimicry is the structural similarity between elongation factor G (EF-G) and the elongation factor Tu – tRNA (EF-Tu: tRNA) complex (Liang and Landweber 2005, and references therein). One possible explanation for this similarity is that both EF-G and EF-Tu : tRNA are bound to the A-site of the ribosome during the elongation stage of protein biosynthesis in prokaryotes and therefore must share similar structures (Watson et al. 2004). One especially interesting observation from our results is that 4 out of the 25 genes with 2 or more fRNA secondary structures are the 4 eukaryotic elongation factors in yeast (2 copies each of eEF1 and eEF2). These 4 genes are also two pairs of duplicate genes produced by the whole genome duplication in yeast. In fact, it is intriguing that the copies of eEF2 seem to have similar fRNA coverage while the copies of eEF1 seem to have undergone asymmetric evolution of fRNA coverage (see Table S10). This observation is even more striking because molecular mimicry is observed between EF-G and EF-Tu: tRNA complex in prokaryotes and EF-G is functionally similar to eEF2 and EF-Tu is functionally similar to eEF1 (Watson et al. 2004). So, it will be interesting to see how gene duplication has affected molecular mimicry in yeast elongation factors and to see if this specific example may help explain the enrichment of genes associated with protein biosynthesis in the set of genes with fRNA secondary structures within their coding regions.

References

Byrne, K.P. and Wolfe, K.H. 2006. "Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast gene after whole-genome duplication" *Genetics* [Epub ahead of print]

Cherry, J.M., Ball C, Weng S, Juvik G, Schmidt R, Adler C, Dunn B, Dwight S, Riles L, Mortimer RK, and Botstein D. 1997. "Genetic and physical maps of *Saccharomyces cerevisiae*" *Nature* **387**: 67073

Conant, G.C., and Wolfe, K.H. 2006. "Functional Partitioning of Yeast Co-Expression Networks after Genome Duplication" *PLoS Biology* **4** (4): e109

Drummond, D.A., Raval, A., and Wilke, C.O. 2006. "A Single Determinant Dominates the Rate of Yeast Protein Evolution" *Molecular Biology and Evolution.* **23**(2): 327-337

Fares, M.A., Byrne, K.P., and Wolfe, K.H. 2006. "Rate Asymmetry After Genome Duplication Causes Substantial Long-Branch Attraction Artifacts in the Phylogeny of *Saccharomyces* Species" *Molecular Biology and Evolution* **23** (2): 245-253

Guan, Y., Dunham, M.J., and Troyanskaya, O.G. 2006. "Functional analysis of gene duplications in S*accharomyces cerevisiae*" *Genetics* [Epub ahead of print]

Hughes, A.L, and Friedman, R. 2005. "Gene Duplication and the Properties of Biological Networks" *Journal of Molecular Evolution* **61**: 758-764

Hurst, L.D. 2005. "Preliminary Assessment of the Impact of MicroRNA-Mediated Regulation on Coding Sequence Evolution in Mammals" *Journal of Molecular Evolution* **63**: 174-182

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., Weber, R.J., Haussler, D. and Kent, W.J. 2003. "The UCSC Genome Browser Database" *Nucleic Acids Research* **31**: 51-54

Kellis, M., Birren, B., and Lander, E. 2004. "Proof and evolutionary analysis of ancient genome duplication in yeast *Saccharomyces cerevisiae*" *Nature* **428**:617-24

Kim, P.M., Lu, L.J., Xia, Y. and Gerstein, M.B. 2006. "Relating Three-Dimensional Structures to Protein Networks Provides Evolutionary Insights" *Science* **314**: 1938-1941

Kim, S-H and Yi, S.V. 2006. "Correlated asymmetry between sequence and functional divergence of duplicate proteins in *Saccharomyces cerevisiae*" *Molecular Biology. and Evolution* **23**: 1068-1075

Liang, H. and Landweber, L.F. 2005. "Molecular mimicry: Quantitative methods to study structural similarity between protein and RNA" *RNA* **11**: 1167-1172

Lin, Y-S., J.K. Byrnes, J-K Hwang, and Li, W-H.. 2006. "Codon-usage bias versus gene conversion in the evolution of yeast duplicate genes" *PNAS* **103**: 14412-14416

Lynch, M., and Conery, J.S. 2000. "The evolutionary fate and consequences of duplicate genes" *Science* **290**: 1151-1155

Lynch, M., and Force, A. 2000. "The probability of duplicate gene preservation by sufunctionalization" *Genetics* **154**: 459-473

Ohno, S. 1970, *Evolution by gene duplication*. Springer-Verlag, Berlin.

Ouellet, D.L., Perron, M.P., Gobell, L-A, Plante, P., and Provost, P. 2006. "MicroRNAs in Gene Regulation: When the Smallest Governs It All" Journal of Biomedicine and Biotechnology. Article ID 69616: 20 pages

Pedersen, J.S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E.S., Kent, J., Miller, W., and Haussler, D. 2006. "Identification and Classification of Conserved RNA Secondary Structures in the Human Genome" *PLoS Computational Biology* **2**: e33

Pickford, A.S., and Cogoni, C. 2003. "RNA-mediated gene silencing" *Cellular and Molecular Life Sciences* **60**: 871-882

R Development Core Team. 2006. *R: A language and environment for statistical computing* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou M., Rosenbloom, K., Clawson, H., Spieth, J., Hiller, L.W., Richards, S., Weinstock, G.M., Wilson, R.K., Gibbs, R.A., Kent, W.J., Miller, W., and Haussler, D. 2005. "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes" *Genome Research* **15**:1034-1050

Sugino, R.P., and Innan, H. 2005. "Estimating the Time to the Whole-Genome Duplication and the Duration of Concerted Evolution via Gene Conversion in Yeast" *Genetics* **171**: 63-69

Teichmann, S.A. and Babu, M.M. 2004. "Gene regulatory network growth by duplication" *Nature Genetics* **36**: 492-496

Thompson, J.D. D.G. Higgins, and Gibson, T.J.. 1994. "CLUSTALW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." *Nucleic Acids Research* **22**: 4673-4680.

Washietl, S., Hofacker, I.L, and Stadler, P.F. 2005. "Fast and reliable prediction of noncoding RNAs" *Proceedings of the National Academy of Sciences, U.S.A*. **102** (7): 2454-2459

Watson, J.D., Baker, T.A., Bell, S.P., Gann, A., Levine, M., and Losick, R. 2004. Molecular Biology of the Gene. *5th edition* Pearson Education, Inc.: San Francisco, CA.

Table S1: Regions of Synteny considered in Initial Analysis

syn42	syn17	syn12	syn2
syn46	**syn32**	**syn7**	**syn37**
syn15	**syn40**	**syn10**	**syn23**
syn38	**syn44**	**syn28**	**syn13**
syn3	**syn30**	**syn41**	**syn1**

Table S2: Genes not available from SGD Gene Table from UCSC Genome Browser

Gene	Source	Chr. Num	Start	Stop	Strand
YNR005C	**SGD Other**	**chr14**	**636928**	**637332**	**-**
YIL170W	**SGD Other**	**chr9**	**19847**	**21220**	+
YDR134C	**SGD Other**	**chr4**	**721064**	**721474**	**-**
YMR245W	**SGD Other**	**chr13**	**758562**	**759182**	+
YCR013C	**SGD Other**	**chr3**	**138396**	**139043**	**-**
YLL016W	**SGD Other**	**chr12**	**112846**	**115992**	+
YDR474C	**SGD_features.tab**	**chr4**	**1407456**	**1410086**	**-**

Table S3: Comparison of EvoFold Predictions from Various Alignments

First Alignment	Second Alignment	Folds in First Alignment	Folds Found in Second Alignment	Folds found with Identical Stem Length
sacKud realign	**sacCas**	**27**	**11**	1
sacKud realign	**sacCas realign**	**27**	**16**	7
sacKud realign	**no gaps**	**27**	**19**	18
sacKud realign	**sacCas no gaps**	**27**	**11**	2
sacKud realign	**sacKud no gaps**	**27**	**19**	18
sacKud realign	**realign**	**27**	**24**	24
realign	**sacCas**	**33**	**11**	1
realign	**sacCas realign**	**33**	**14**	4
realign	**sacCas no gaps**	**33**	**11**	2
realign	**no gaps**	**33**	**19**	18
realign	**sacCas sacKud no gaps**	**33**	**11**	2
realign	**sacKud no gaps**	**33**	**19**	18

sacKud no gaps	**sacCas**	**30**	**18**	4
sacKud no gaps	**casCas** **realign**	**30**	**10**	0
sacKud no gaps	**sacCas** **no gaps**	**30**	**18**	8
sacKud no gaps	**no gaps**	**30**	**30**	30
sacKud no gaps	**sacCas** **sacKud** **no gaps**	**30**	**18**	8
sacCas sacKud no gaps	**sacCas**	**39**	**39**	35
sacCas sacKud no gaps	**sacCas** **realign**	**39**	**30**	25
sacCas sacKud no gaps	**sacCas** **no gaps**	**39**	**39**	39
sacCas sacKud no gaps	**no gaps**	**39**	**18**	8
no gaps	**sacCas**	**30**	**18**	4
no gaps	**sacCas** **realign**	**30**	**10**	0
no gaps	**sacCas** **no gaps**	**30**	**18**	8
sacCas no gaps	**sacCas**	**39**	**39**	35
sacCas no gaps	**sacCas** **realign**	**39**	**20**	25
sacCas **realign**	sacCas	43	30	25

The term "align" refers to alignments that were realigned with CLUSTALW. The term "no gaps" refers to alignments where any gaps present in all the species being compared have been removed (this is due gaps that were created by species from the 7 species comparison that are not included in a particular alignment). The abbreviations for various species are as follows: sacCer = *Saccharomyces cerevisiae*; sacPar = *Saccharomyces paradoxus*,;sacMik = *Saccharomyces mikatae*; sacBay = *Saccharomyces baynus*; sacCas = *Saccharomyces castelii*; sacKud = *Saccharomyces kudriavzevii*. All alignments include sacCer, sacPar, sacMik, and sacBay, so they are not explicitly listed above. All folds were predicted from a region on Chromosome 3 from 208077 to 289301 and on chromosome 14 from 664217 to 715435 (UCSC coordinates).

Table S4: Recovered Annotations using Various Prediction Methods

Prediction Category	tRNA (out of 275)	rRNA (out of 11)	snoRNA (out of 66)	snRNA (out of 6)	Misc RNA (out of 14)
EvoFold 4 all	34	7	21	4	11
EvoFold 4 FPS > 0	25	5	14	2	6
EvoFold 5 all	63	7	28	5	11
EvoFold 5 FPS > 0	45	6	22	5	11
EvoFold 6 all	85	9	18	5	8
EvoFold 6 FPS > 0	77	13	2	8	6
RNAz 4 p>0.5	184	8	48	6	14
RNAz 4 p>0.9	112	7	32	3	12
RNAz 5 p>0.5	163	7	45	5	14
RNAz 5 p>0.9	119	7	27	3	13
RNAz 6 p> 0.5	150	7	31	4	14
RNAz 6 p>0.9	107	7	25	4	12

Number in 1st column (e.g. 4, 5, or 6) refers to the number of species used in multi-species alignment. FPS stands for folding potential score.

Table S5: Table 1: Pearson Correlation Coefficients (ρ) for All Folds

Prediction Category	SRR~SRA	SRR~SRK	SRR~SRF	SRR~SRN
EvoFold 4 all	-0.001	-0.138***	-0.031	0.011
EvoFold 4 FPS > 0	-0.013	-0.113**	-0.027	-0.054
EvoFold 5 all	0.005	-0.069	-0.030	-0.060
EvoFold 5 FPS > 0	-0.015	-0.052	-0.016	-0.024
EvoFold 6 all	-0.068	-0.048	-0.010	-0.032
EvoFold 6 FPS > 0	-0.030	-0.032	0.032	-0.018
RNAz 4 p>0.5	0.090*	0.062	-0.128***	-0.184****
RNAz 4 p>0.9	0.084	0.109**	-0.084*	-0.120**
RNAz 5 p>0.5	0.068	0.079*	-0.099**	-0.248****
RNAz 5 p>0.9	0.016	0.070	-0.121**	-0.186****
RNAz 6 p> 0.5	-0.014	0.0928**	-0.109**	-0.184****
RNAz 6 p>0.9	-0.094*	0.031	-0.083*	-0.101**

Significance levels: *=p<0.10, ** = p<0.05, ***=p<0.01, ****p<0.001

Table S6: Table 1: Pearson Correlation Coefficients (ρ) for Coding Folds

Prediction Category	SRR~SRA	SRR~SRK	SRR~SRF	SRR~SRN
EvoFold 4 all	0.052	-0.179*	-0.103	-0.069
EvoFold 4 FPS > 0	0.027	-0.149	-0.129	-0.161
EvoFold 5 all	0.031	-0.130	-0.024	-0.144*
EvoFold 5 FPS > 0	0.082	-0.113	0.031	-0.157
EvoFold 6 all	-0.056	-0.033	-0.066	-0.144*
EvoFold 6 FPS > 0	0.046	-0.006	-0.014	-0.152
RNAz 4 p>0.5	0.112*	0.102*	-0.145***	-0.235****
RNAz 4 p>0.9	0.167**	0.195***	-0.133*	-0.246****
RNAz 5 p>0.5	0.095	0.125**	-0.134**	-0.328****
RNAz 5 p>0.9	0.087	0.099	-0.194**	-0.313****
RNAz 6 p> 0.5	0.029	0.115*	-0.138**	-0.274****
RNAz 6 p>0.9	-0.152	0.001	-0.099	-0.192**

Significance levels: *=p<0.10, ** = p<0.05, ***=p<0.01, ****p<0.001

Table S7: Table 1: Pearson Correlation Coefficients (ρ) for 5` Flank Folds

Prediction Category	SRR~SRA	SRR~SRK	SRR~SRF	SRR~SRN
EvoFold 4 all	0.074	-0.310**	-0.011	0.110
EvoFold 4 FPS > 0	0.185	-0.269	-0.174	0.106
EvoFold 5 all	0.089	-0.132	-0.084	-0.017
EvoFold 5 FPS > 0	0.061	-0.087	-0.232	0.099
EvoFold 6 all	0.089	-0.320**	-0.061	0.184
EvoFold 6 FPS > 0	0.282	-0.338	-0.128	0.186
RNAz 4 p>0.5	0.077	0.040	-0.097	-0.058
RNAz 4 p>0.9	0.097	0.067	0.061	-0.170
RNAz 5 p>0.5	-0.098	0.025	-0.116	-0.061
RNAz 5 p>0.9	-0.246	0.086	-0.124	-0.027
RNAz 6 p> 0.5	-0.193	0.142	-0.057	0.126
RNAz 6 p>0.9	-0.134	0.237	-0.183	0.068

Significance levels: *=p<0.10, ** = p<0.05, ***=p<0.01, ****p<0.001

Table S8: Table 1: Pearson Correlation Coefficients (ρ) for 3` Flank Folds

Prediction Category	SRR~SRA	SRR~SRK	SRR~SRF	SRR~SRN
EvoFold 4 all	-0.056	-0.052	0.045	-0.089
EvoFold 4 FPS > 0	-0.125	-0.170	0.075	-0.106
EvoFold 5 all	-0.024	0.119	-0.004	-0.064
EvoFold 5 FPS > 0	-0.080	0.008	0.038	-0.027
EvoFold 6 all	-0.170	0.016	0.131	-0.106
EvoFold 6 FPS > 0	-0.332*	-0.202	0.320*	0.0309
RNAz 4 p>0.5	0.085	-0.251**	-0.071	-0.168
RNAz 4 p>0.9	-0.204	-0.126	-0.173	0.186
RNAz 5 p>0.5	0.838	0.333	0.171	-0.434
RNAz 5 p>0.9	-0.041	0.189	-0.184	-0.027

RNAz 6 p> 0.5	-0.061	0.057	-0.104	-0.106
RNAz 6 p>0.9	-0.061	0.115	-0.053	0.019

Significance levels: *=p<0.10, ** = p<0.05, ***=p<0.01, ****p<0.001

Table S9: Table 1: Pearson Correlation Coefficients (ρ) for Intron Folds

Prediction Category	SRR~SRA	SRR~SRK	SRR~SRF	SRR~SRN
EvoFold 4 all	-0.110	-0.197	-0.330	-0.370
EvoFold 4 FPS > 0	0.645	0.075	0.220	-0.447
EvoFold 5 all	0.069	-0.058	-0.474	-0.220
EvoFold 5 FPS > 0	0.495	0.315	-0.369	-0.514
EvoFold 6 all	0.642	0.214	0.705	-0.222
EvoFold 6 FPS > 0	0.615	0.099	0.688	0.000
RNAz 4 p>0.5	0.838	0.333	0.172	-0.434
RNAz 4 p>0.9	N/A	N/A	N/A	N/A
RNAz 5 p>0.5	0.838	0.333	0.172	-0.434
RNAz 5 p>0.9	N/A	N/A	N/A	N/A
RNAz 6 p> 0.5	N/A	N/A	N/A	N/A
RNAz 6 p>0.9	N/A	N/A	N/A	N/A

Significance levels: *=p<0.10, ** = p<0.05, ***=p<0.01, ****p<0.001. When "N/A" appears in a column, this means that either SD=0 or there are not enough observations.

Table S10: Impact of fRNAs on Eukaryotic Elongation Factors

Gene	# fRNAs	% Coverage	dN
YPR080W (TEF1/eEF1)	5	15.3	0.016
YBR118W (TEF2/eEF1)	2	3.99	N/A
YOR133W (EFT1/eEF2)	2	2.21	N/A
YDR385W (EFT2/eEF2)	2	2.21	0.257

Note: dN is not currently available because I am using values calculated from Wall et al. 2005. I will calculate these values myself within the next couple weeks

Table S11: Unreliable Protein-coding gene annotations

Name	Type	Details
YDL133W	Uncharacterized	Hypothetical protein
YGR053C	Uncharacterized	Hypothetical protein
YIL041W	Uncharacterized	Golgi vesicle protein of unknown function; localizes to both early and late Golgi vesicles
YMR102C	Uncharacterized	Protein of unknown function; transcription is activated by paralogous transcription factors Yrm1p and Yrr1p along with genes involved in multidrug resistance; mutant shows increased resistance to azoles; YMR102C is not an essential gene
YLR281C	Uncharacterized	Putative protein of unknown function; green fluorescent protein (GFP)-fusion protein localizes to mitochondria; YLR281C is not an essential gene
YLR156C-A	Uncharacterized	Putative protein of unknown function identified by fungal homology comparisons and RT-PCR; this ORF partially overlaps RND5-3
YOR396W	Uncharacterized	Hypothetical protein
YLR154C-H	Uncharacterized	Putative protein of unknown function identified by fungal

		homology comparisons and RT-PCR; this ORF partially overlaps RND5-2
YAR028W	Uncharacterized	Putative integral membrane protein, member of DUP240 gene family
YKL100C	Uncharacterized	Hypothetical protein
YPL009C	Uncharacterized	Hypothetical protein
YDR475C	Uncharacterized	Protein of unknown function; previously annotated as two separate ORFs, YDR474C and YDR475C, which were merged as a result of corrections to the systematic reference sequence
YPR159C-A	Uncharacterized	Identified by gene-trapping, microarray-based expression analysis, and genome-wide homology searching
YLR154C-G	Uncharacterized	Putative protein of unknown function identified by fungal homology comparisons and RT-PCR; this ORF is contained within RDN25-2 and RDN37-2
YMR013W-A	Dubious	Dubious open reading frame unlikely to encode a protein, based on available experimental and comparative sequence data; completely overlaps the characterized snoRNA gene snR73
YNL050C	Uncharacterized	Putative protein of unknown function; YNL050c is not an essential gene
YBR281C	Uncharacterized	Putative protein of unknown function; green fluorescent protein (GFP)-fusion protein localizes to the cytoplasm and nucleus; YBR281C is not an essential gene
YLR108C	Uncharacterized	Protein of unknown function; green fluorescent protein (GFP)-fusion protein localizes to the nucleus; YLR108C is not an esssential gene
YBR241C	Uncharacterized	Putative protein of unknown function; green fluorescent protein (GFP)-fusion protein localizes to the vacuolar membrane; YBR241C is not an essential gene
YDL038C	Uncharacterized	Putative protein of unknown function; YDL038C is not an essential gene
YBR094W	Uncharacterized	Hypothetical protein
YKR075C	Uncharacterized	Protein of unknown function; similar to YOR062Cp and Reg1p; expression regulated by glucose and Rgt1p
YOR111W	Uncharacterized	Hypothetical protein
YNL042W-B	Uncharacterized	Putative protein of unknown function
YOL131W	Uncharacterized	Hypothetical protein
YOL013W-B	Dubious	Similar to probable membrane protein YLR334C and ORF YOL106W
YPR108W-A	Uncharacterized	Identified by fungal homology and RT-PCR
YHR131C	Uncharacterized	Putative protein of unknown function; green fluorescent protein (GFP)-fusion protein localizes to the cytoplasm
YLR089C	Uncharacterized	Putative alanine transaminase (glutamic pyruvic transaminase)
YJR112W-A	Uncharacterized	Putative protein of unknown function; identified based on homology to Ashbya gossypii
YGL117W	Uncharacterized	Hypothetical protein
YBL029C-A	Uncharacterized	Protein of unknown function; green fluorescent protein (GFP)-fusion protein localizes to the cell periphery; has potential orthologs in Saccharomyces species and in Yarrowia lipolytica
YOR291W	Uncharacterized	Hypothetical protein
YMR233W	Uncharacterized	Protein of unknown function with similarity to components of human SWI/SNF complex including SMRD3; green

		fluorescent protein (GFP)-fusion protein localizes to the cytoplasm, nucleus and nucleolus; YMR233W is not an essential gene
YOR072W-B	Uncharacterized	Identified by expression profiling and mass spectrometry
YLR099W-A	Uncharacterized	Putative protein of unknown function
YGR130C	Uncharacterized	Putative protein of unknown function; green fluorescent protein (GFP)-fusion protein localizes to the cytoplasm; specifically phosphorylated in vitro by mammalian diphosphoinositol pentakisphosphate (IP7)
YLR154W-E	Dubious	Dubious open reading frame unlikely to encode a protein; encoded within the the 35S rRNA gene on the opposite strand
YAR031W	Verified	Really bad multiz alignment
YBL046W	Verified	Hypothetical ORF (with premature stop codons)

Figure S1: SR Correlations for All Folds in Stringent Dataset

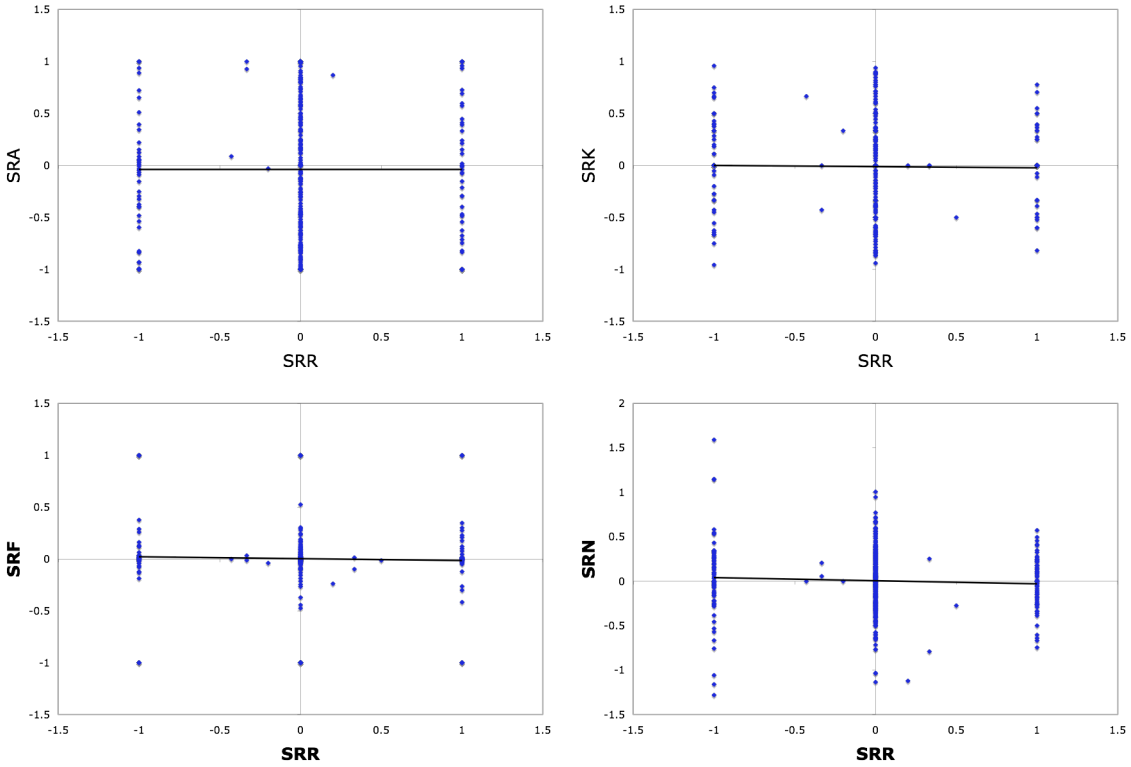


Figure S2: SR Correlations for Coding Folds in Stringent Dataset

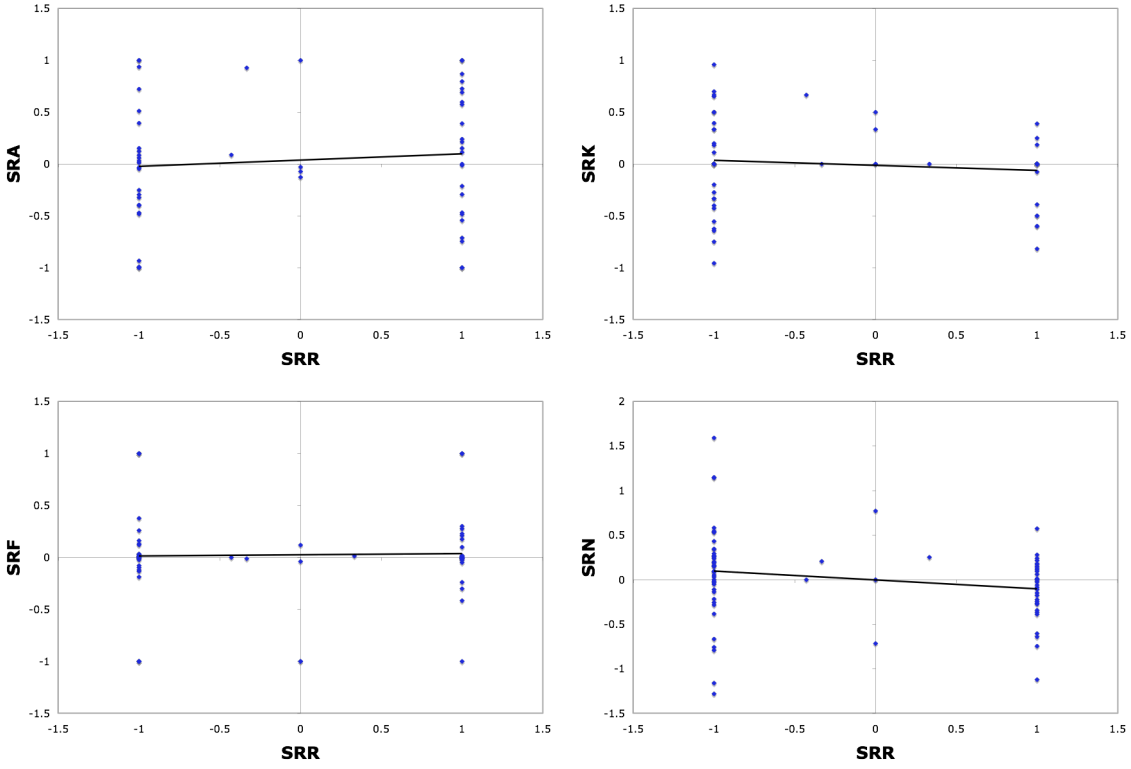


Figure S3: SR Correlations for 5` Folds in Stringent Dataset

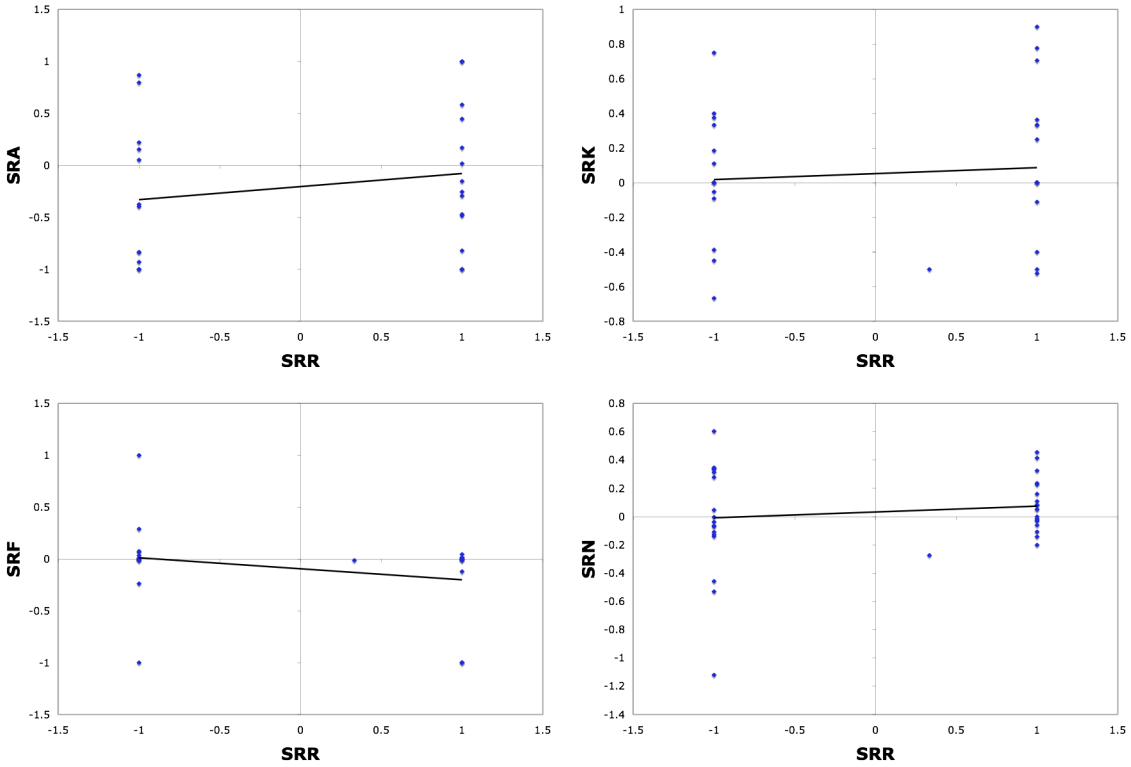


Figure S4: SR Correlation for 3` Folds in Stringent Dataset

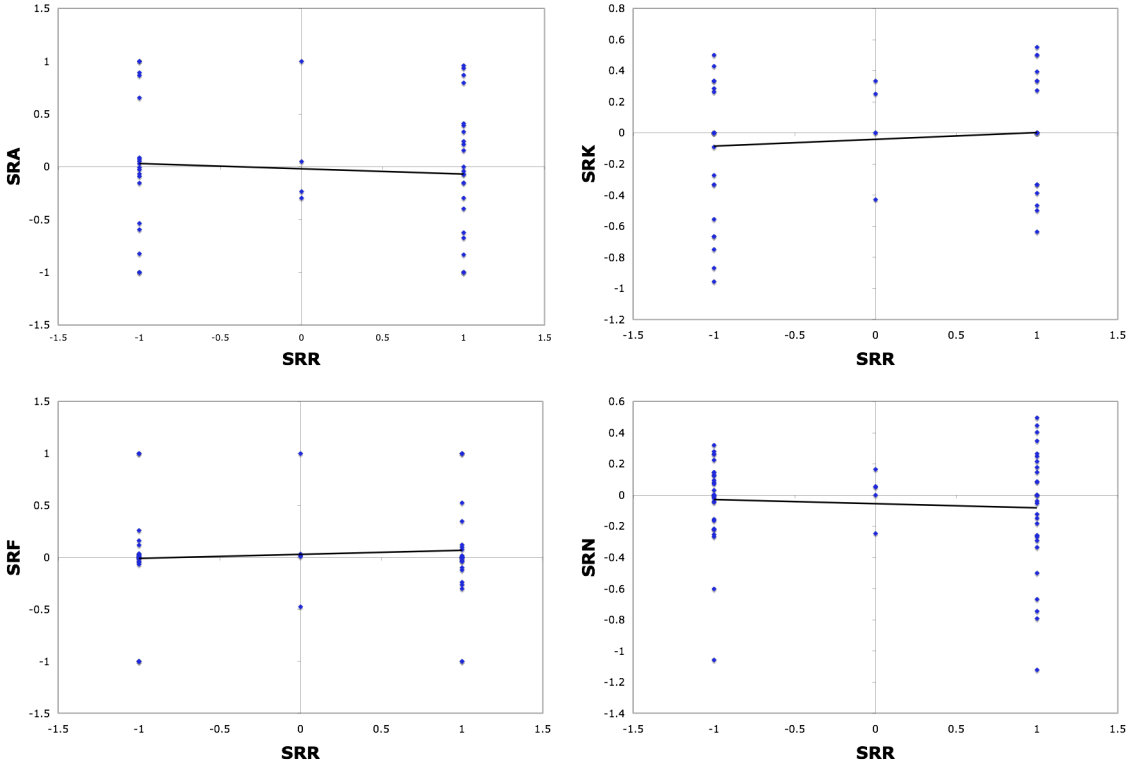


Figure S5: SR Correlations for Intron Folds in Stringent Dataset

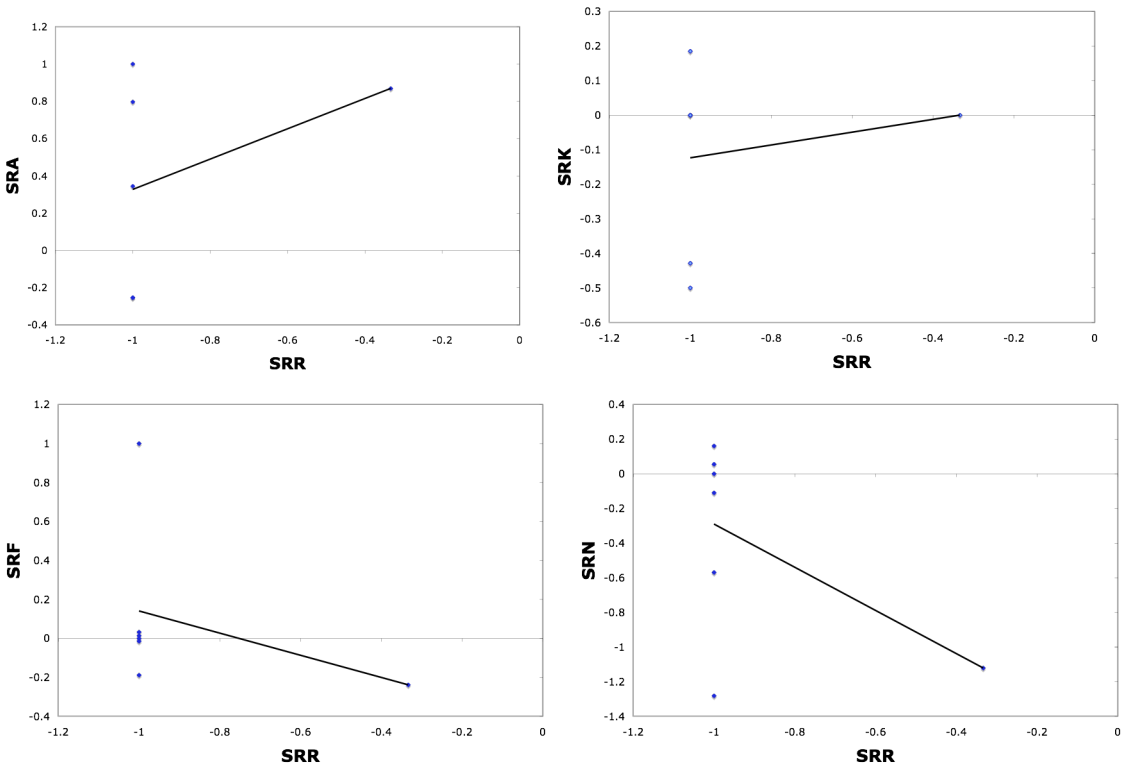


Figure S6: Correlation between FPS and Length for 4-species comparison

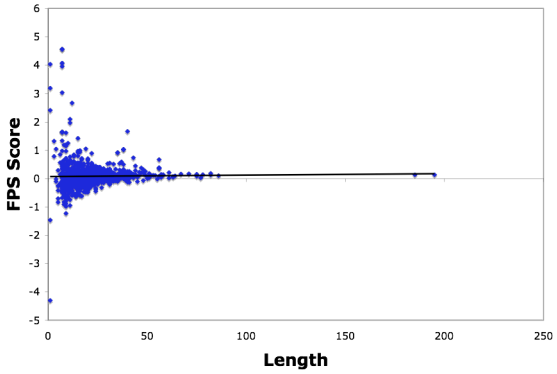


Figure S7: Correlation between FPS and Length for 5-species comparisons

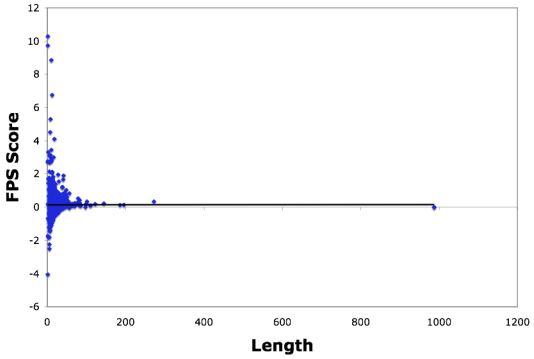


Figure S8: Correlation between FPS and Length for 6-species comparisons

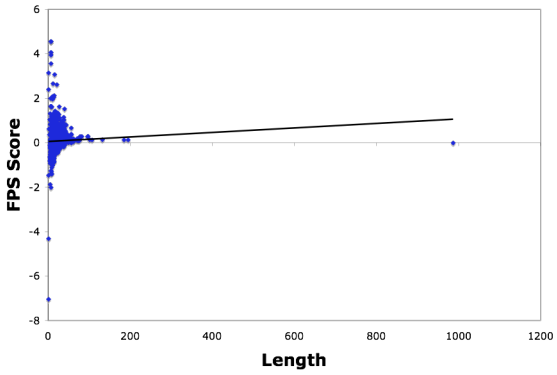
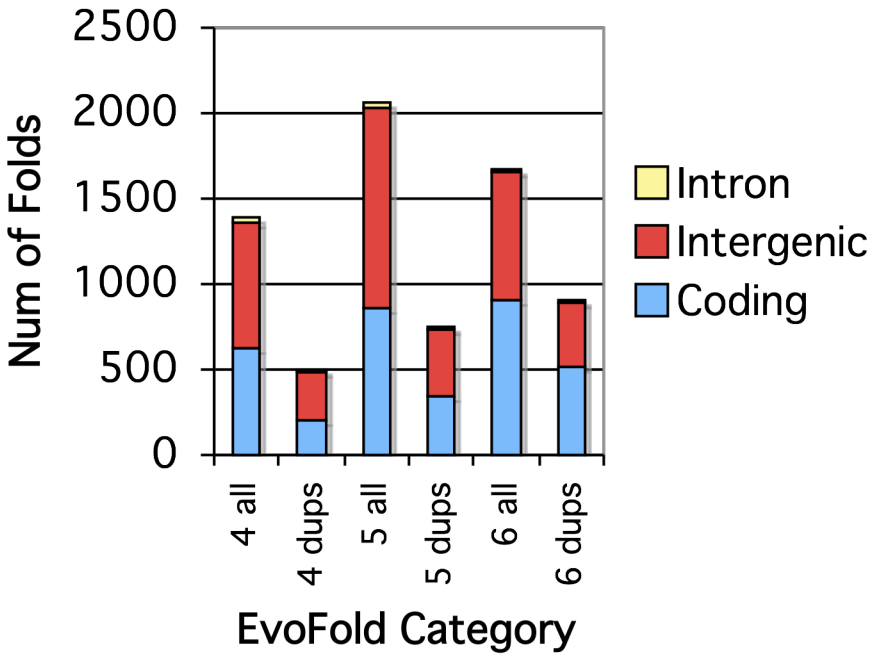
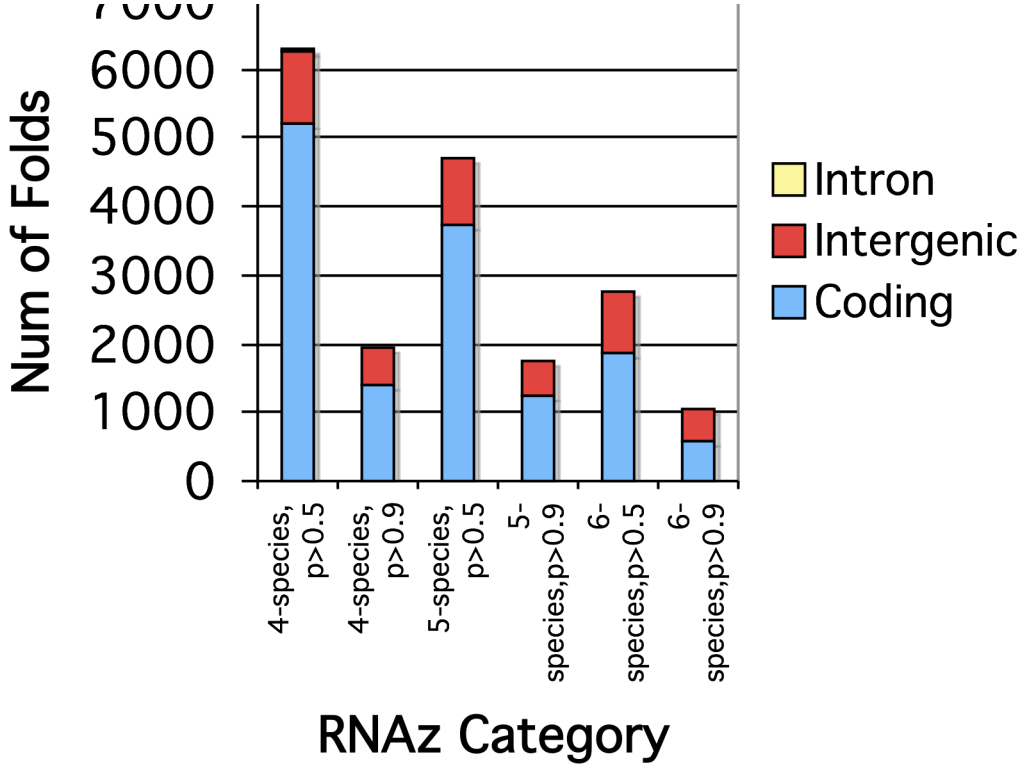


Figure S9: Categories of EvoFold predictions for various comparisons

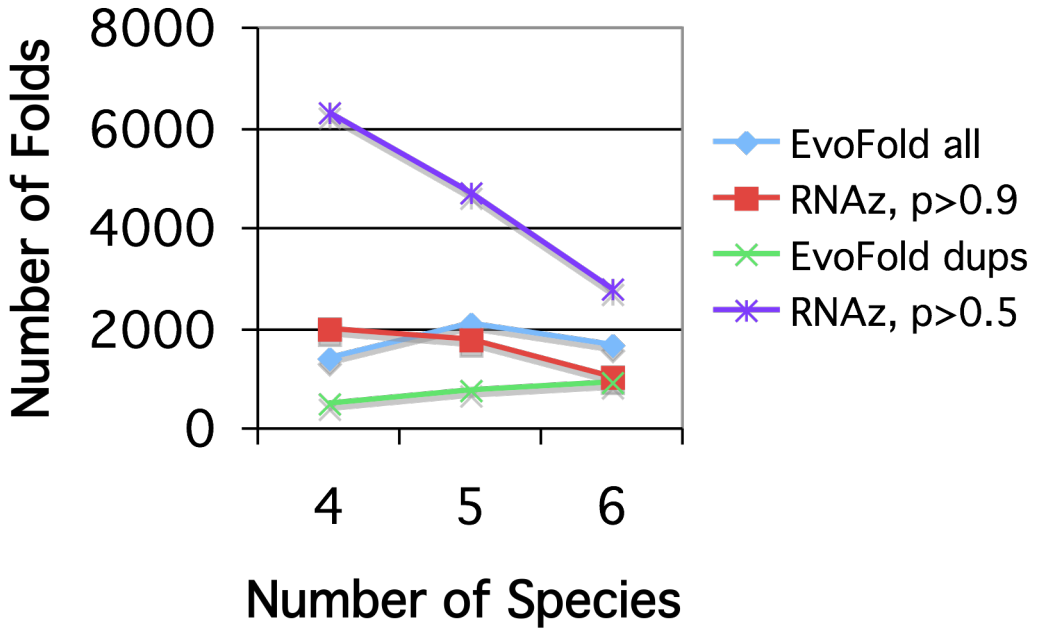


Note: "all" refers to EvoFold predictions from phastCons blocks throughout the yeast genome, "dups" refers to Multiz alignment for WGD pairs (including flanking regions)

Figure S10: Categories of RNAz predictions for various comparisons

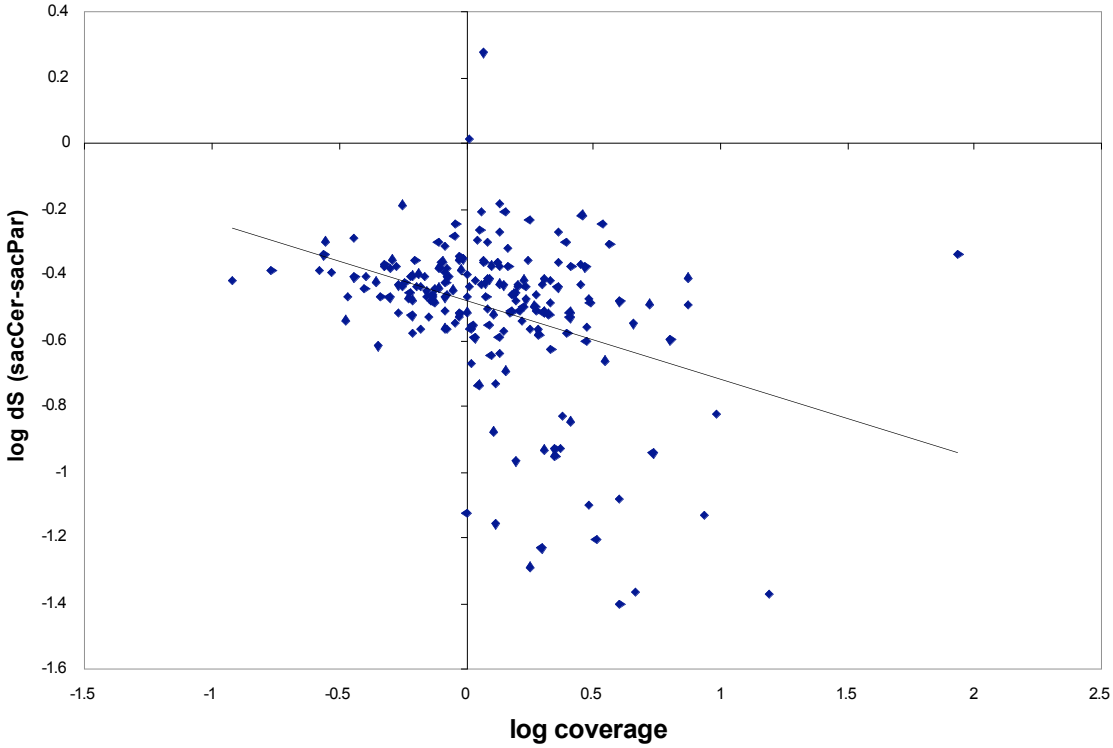


S11: Total Number of Folds Predicted by Various Methods



Note: "all" refers to EvoFold predictions from phastCons blocks throughout the yeast genome, "dups" refers to Multiz alignment for WGD pairs (including flanking regions)

S12: Constraint of Synonymous Site Evolution Due to the Presence of Coding fRNAs



(ρ= -0.357, p-value = 3.50 x 10^{-7})